МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)

ИНСТИТУТ ЛИНГВИСТИКИ
Учебно-научный центр компьютерной лингвистики

Алгоритмы машинного обучения

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

45.04.03 Фундаментальная и прикладная лингвистика
Код и наименование направления подготовки/специальности

Магистерская программа: Фундаментальная и компьютерная лингвистика
Наименование направленности (профиля)/ специализации

Уровень высшего образования: магистратура

Форма обучения: очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2024

Алгоритмы машинного обучения
Рабочая программа дисциплины

Составитель(и):

старший преподаватель А.М.Ивойлова

Ответственный редактор:

к.ф.н, доцент Н.А.Коротаев

УТВЕРЖДЕНО
Протокол заседания УНЦ компьютерной лингвистики
№ 5 от 26 марта 2024 г.

ОГЛАВЛЕНИЕ

1. Пояснительная записка.....
1.1. Цель и задачи дисциплины.....
1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций.....
1.3. Место дисциплины в структуре образовательной программы.....
2. Структура дисциплины.....
3. Содержание дисциплины.....
4. Образовательные технологии.....
5. Оценка планируемых результатов обучения.....
5.1 Система оценивания.....
5.2 Критерии выставления оценки по дисциплине.....
5.3 Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине.....
6. Учебно-методическое и информационное обеспечение дисциплины.....
6.1 Список источников и литературы.....
6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет».....
6.3 Профессиональные базы данных и информационно-справочные системы.....
7. Материально-техническое обеспечение дисциплины.....
8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов.....
9. Методические материалы.....
9.1 Планы семинарских/ практических/ лабораторных занятий.....
9.2 Методические рекомендации по подготовке письменных работ.....
9.3 Иные материалы.....

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Цель курса — ознакомить студентов с такой областью знаний, как искусственный интеллект, дать представление о наиболее широко известных и распространенных классических алгоритмах машинного обучения, основных его понятиях и методах.

Задачи дисциплины: научить студентов выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий с помощью классических методов машинного обучения.

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.1 Знает принципы сбора, отбора и обобщения информации	Знать: – основные типы формальных лингвистических моделей, принципы применения математического аппарата для формализации языковых явлений; Уметь: – пользоваться основными методами, способами и средствами получения, хранения, переработки информации; – пользоваться лингвистически ориентированными программными продуктами; Владеть: – принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и иных баз данных и баз знаний).
ПК-3 Способен использовать лингвистические технологии для проектирования систем автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистических компонентов интеллектуальных и	ПК-3.1 Знает основные системы автоматической обработки звучащей речи и текстов на естественном языке; базовые принципы автоматической обработки языковых данных; основные интеллектуальные и информационные электронные системы и принципы работы с ними	Знать: – принципы работы лингвистически ориентированных программных продуктов; Уметь: – пользоваться лингвистически ориентированными программными продуктами; Владеть: – навыками использования лингвистически ориентированных программных продуктов.

информационных электронных систем		

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Алгоритмы машинного обучения» является элективной дисциплиной и относится к части, формируемой участниками образовательных отношений блока дисциплин учебного плана.

Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин и прохождения практик: Основания математики, Основы языка программирования Python.

В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения следующих дисциплин и прохождения практик: Основы глубинного обучения, Научно-исследовательская работа, Преддипломная практика.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 108 академических часов.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
2	Практические занятия	30
2	Экзамен	18
	Всего:	48

Объем дисциплины (модуля) в форме самостоятельной работы обучающихся составляет 60 академических часов.

3. Содержание дисциплины

№	Наименование раздела дисциплины	Содержание
1.	Введение. Основные понятия машинного обучения.	Знакомство с основными понятиями машинного обучения, такими, как выборка, обучающая и валидационная выборка, целевая переменная, метрики, обучение, кросс-валидация.
2.	Алгоритмы регрессии. Линейная регрессия.	Задача регрессии. Алгоритм линейной регрессии. Методы градиентного спуска. SGD, MBGD.
3.	Нормализация и масштабирование.	L1 и L2 нормализация. Lasso, Ridge. Масштабирование данных.
4.	Линейные алгоритмы классификации.	Задача классификации. Алгоритмы: Логистическая регрессия, Метод опорных векторов.
5.	Нелинейные алгоритмы классификации.	Задача классификации. Алгоритмы: Наивный Байес, Метод K ближайших соседей.
6.	Решающие деревья.	Деревья решений для решения задач классификации и регрессии. Переобучение. Random Forest.
7.	Ансамбли.	Смешивание алгоритмов. Stacking, Bagging, Boosting.
8.	Работа с текстами.	Понятие эмбеддингов. Применение алгоритмов

		машинного обучения к текстовым данным.
9.	Задача кластеризации.	Задача кластеризации. Алгоритмы K-means, DBSCAN, PCA, t-SNE. Снижение размерности.

4. Образовательные технологии

Для проведения учебных занятий по дисциплине используются различные образовательные технологии. Для организации учебного процесса может быть использовано электронное обучение и (или) дистанционные образовательные технологии.

5. Оценка планируемых результатов обучения

5.1 ***Система оценивания***

Форма контроля	**Макс. количество баллов**	
	За одну работу	**Всего**
Текущий контроль:		
- домашние задания	*5 баллов*	*30 баллов*
- выполнение заданий на семинаре	*5 баллов*	*10 баллов*
- участие в соревновании	*20 баллов*	*20 баллов*
Промежуточная аттестация – зачет		40 баллов
Итого за семестр		100 баллов

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	Традиционная шкала		Шкала ECTS
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

5.2 Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	отлично/ зачтено	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне –

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		«высокий».
82-68/ C	хорошо/ зачтено	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	удовлетво-рительно/ зачтено	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	неудовлет-ворительно/ не зачтено	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3 Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

1. Обучение алгоритма регрессии на датасете House sales (https://www.kaggle.com/datasets/harlfoxem/housesalesprediction)
2. Обучение алгоритма классификации на датасете Students (https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams)
3. Обучение алгоритма классификации на датасете Twitter Sentiment (https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset)
4. Участие в соревновании New York Taxi Fare Prediction (https://www.kaggle.com/competitions/new-york-city-taxi-fare-prediction)

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 Список источников и литературы

1. Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — Москва : Издательство Юрайт, 2022. — 85 с. — (Высшее образование). — ISBN 978-5-534-

15561-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/508804 (дата обращения: 31.10.2022).
2. Мюллер А., Гвидо. С. Введение в машинное обучение с помощью Python. М. 2016-2017. URL: https://edu.vsu.ru/pluginfile.php/1246708/mod_resource/content/1/1myuller_a_gido_s_vvedenie_v_mashinnoe_obuchenie_s_pomoshch_y.pdf
3. Рашка С. Python и машинное обучение. М: ДМК Пресс, 2017.

6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. https://scikit-learn.org/stable/
2. https://www.kaggle.com/
3. http://www.machinelearning.ru/
4. nlp-progress.com
5. Национальная электронная библиотека (НЭБ) www.rusneb.ru
6. ELibrary.ru Научная электронная библиотека www.elibrary.ru
7. Электронная библиотека Grebennikon.ru www.grebennikon.ru
8. Cambridge University Press
9. ProQuest Dissertation & Theses Global
10. SAGE Journals
11. Taylor and Francis
12. JSTOR

6.3 Профессиональные базы данных и информационно-справочные системы

Доступ к профессиональным базам данных: https://liber.rsuh.ru/ru/bases

База данных для машинного обучения:
www.kaggle.com

7. Материально-техническое обеспечение дисциплины

Для обеспечения дисциплины используется материально-техническая база образовательного учреждения: учебные аудитории, оснащённые компьютером и проектором для демонстрации учебных материалов.

Состав программного обеспечения:
1. Windows
2. Microsoft Office
3. Kaspersky Endpoint Security
4. **Python 3.9**
5. **Visual Studio Code**
6. **PyCharm Community Edition**
7. **Jupyter Lab**

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением или могут быть заменены устным ответом; обеспечивается индивидуальное равномерное освещение не менее 300 люкс; для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств; письменные задания оформляются увеличенным шрифтом; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.
- для глухих и слабослышащих: лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования; письменные задания выполняются на компьютере в письменной форме; экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.
- для лиц с нарушениями опорно-двигательного аппарата: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих: в печатной форме увеличенным шрифтом, в форме электронного документа, в форме аудиофайла.
- для глухих и слабослышащих: в печатной форме, в форме электронного документа.
- для обучающихся с нарушениями опорно-двигательного аппарата: в печатной форме, в форме электронного документа, в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих: устройством для сканирования и чтения с камерой SARA CE; дисплеем Брайля PAC Mate 20; принтером Брайля EmBraille ViewPlus;
- для глухих и слабослышащих: автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих; акустический усилитель и колонки;
- для обучающихся с нарушениями опорно-двигательного аппарата: передвижными, регулируемыми эргономическими партами СИ-1; компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1 Планы семинарских/ практических/ лабораторных занятий

Тема 1. Введение в машинное обучение. Основные понятия. Алгоритмы регрессии. Линейная регрессия.

Знакомство с библиотекой scikit-learn. Обучение алгоритма линейной регрессии.

Задания:

1. Самостоятельно обучить алгоритм линейной регрессии для предложенного датасета.
2. Проанализировать признаки предложенного датасета. Попытаться добиться улучшения качества работы алгоритма.

Тема 2. Градиентный спуск. Линейная регрессия. Работа с признаками.

Обучение алгоритма линейной регрессии. Работа с признаками в обучающих данных.

Задания:

1. Самостоятельно обучить алгоритм линейной регрессии для предложенного датасета.
2. Проанализировать признаки предложенного датасета. Попытаться добиться улучшения качества работы алгоритма.

Тема 3. Методы нормализации. Масштабирование признаков.

Знакомство с L1 и L2 нормализацией; обучение алгоритмов Lasso, Ridge, ElasticNet. Применение видов масштабирования: StandardScaler, MinMaxScaler.

Задания:

1. Самостоятельно обучить алгоритм линейной регрессии с нормализацией для предложенного датасета.
2. Проанализировать признаки предложенного датасета. Попытаться добиться улучшения качества работы алгоритма.

Тема 4. Алгоритмы линейной классификации.

Знакомство с алгоритмами Logistic Regression, SVM

1. Обучить алгоритмы классификации на датасете “фотографии президентов США”. Оценить, с каким качеством работают эти алгоритмы.
2. Попробовать добиться лучшего качества работы алгоритмов.

Тема 5. Нелинейные алгоритмы классификации. Решающие деревья.

Знакомство с решающими деревьями. Случайный лес для задач классификации и регрессии.

Задания:

1. Обучить решающее дерево на предложенном датасете. Оценить степень переобучения.
2. Попробовать снизить переобучение указанными методами.

Тема 6. Работа с текстами.

Применение изученных алгоритмов МО к текстовым данным.

Задания:

1. Обучить алгоритм классификации на предложенном датасете. Поработать с его признаками.

Тема 7. Задача кластеризации.

Алгоритмы кластеризации.

Задания:

1. Обучить алгоритм кластеризации на предложенном датасете.
2. Попробовать использовать результаты получившейся модели в качестве дополнительных признаков при обучении классификации.

9.2 Иные материалы

Все необходимые для обучения материалы публикуются по адресу https://github.com/rsuh-python/ в соответствующих репозиториях.